Premiers résultats de l'étude Webspam/Qualité des frères Peyronnet

Par Guillaume et Sylvain Peyronnet

Domaine :	Recherche	Référencement
Niveau :	Pour tous	Avancé

Depuis de nombreux mois, plusieurs chercheurs, dont les auteurs de cet article, travaillent sur une étude permettant d'obtenir des indications claires sur la notion de "contenu de bonne qualité" et, par incidence, sur ce que les moteurs considèrent comme étant du spam sur le Web. Vous trouverez donc ici les premiers résultats, parfois surprenants, de cette analyse sur 5 critères de pertinence importants pour Google et consorts.

Dans cet article nous allons présenter les résultats préliminaires de ce que la communauté SEO a appellé « l'étude webspam des frères Peyronnet ». En effet, il y a maintenant deux ans, alors que nous mettions en place des formations présentant les algorithmes utilisés par les moteurs de recherches pour détecter les contenus « spammy », nous nous sommes posé la question de savoir si les critères utilisés par Ntoulas et ses coauteurs en 2006 (voir notre article en juin 2014 : http://recherche-referencement.abondance.com/2014/06/spam-de-contenu-comment-le-reconnaitre.html) étaient toujours valides. Il semblait évident que cela n'était pas le cas, et nous avons décidé de mettre en place notre propre étude, pour laquelle la communauté SEO francophone nous a aidé, en donnant de l'argent lors d'un crowdfunding, et du temps lors de l'étape de notation.

Notre équipe, nos objectifs

Cette étude, c'est d'abord le travail d'une équipe. Les trois personnes impliqués de manière continue sont Thomas Largillier (maitre de conférences à l'Université de Caen Basse-Normandie), Guillaume Peyronnet (Nalrem Médias) et Sylvain Peyronnet (ix-labs et Qwant), auteur de cet article. Nous avons été ponctuellement aidé par d'autres personnes, comme par exemple Laurent Bourrely pour le crowdfunding et la détermination des critères à étudier.

Les objectifs de ce travail sont multiples, mais le principal est de déterminer les critères utilisés par les internautes (de manière généralement implicites) pour décider de la qualité d'une page web. On note ici qu'on parle bien de qualité, ce qui est un problème plus complexe que simplement déterminer si une page est du spam ou non. De cet objectif principal découlent des sous-objectifs : avoir un *dataset* pour la communauté scientifique, obtenir un classifieur (par exemple via

l'algorithme C4.5 [2]) pour les moteurs et les référenceurs, etc. Nous évoquerons ces objectifs plus loin dans l'article.

La méthodologie

Acquisition des données

Nous avons tout d'abord effectué un crawl de pages web que nous avons stockées. Ce crawl a été effectué à la fin de l'année 2014, et a utilisé comme noyau les pages web correspondant aux SERP de Qwant pour les 10 000 requêtes les plus demandées au moteur. Nous avons ainsi obtenu environ 155 000 pages web.

Notation

Nous avons ensuite demandé à des volontaires de noter ces pages web. Pour cela nous avons ouvert un site dédié à la notation, dont la figure 1 présente une capture de l'interface.

Le point le plus important est de bien comprendre ce que l'on a demandé comme action aux volontaires. Pour chaque page du dataset, nous attendions des notations par 3 à 5 volontaires (selon le degré d'accord entre les volontaires). Il était demandé de noter chaque page comme :

- « Pas en Français » : la page est dans une autre langue.
- « Je ne sais pas » : page d'erreur ou page impossible à noter selon le volontaire.
- « Haute qualité » le contenu paraît tel qu'il semble normal qu'il soit poussé par un moteur de recherche.
- « Basse qualité » : Cette page peut apparaître au sein des SERP, mais pas de manière générale, seulement sur des requêtes spécifiques. Il s'agit typiquement de pages qui présentent mal des informations pertinentes, ou qui sont lacunaires.
- « SPAM » : « Je pense que cette page ne mérite pas d'être proposée par un moteur de recherches, quelle que soit la requête (ou je vois des manipulations grossières) ».

Fig. 1. Le site web mis en place pour l'étude.

Une fois les notations obtenues, nous les avons retravaillées. Plus précisément, on a associé une valeur à chaque notation (spam = 1 point, basse qualité = 3 point et haute qualité = 5 point). Chaque page a eu par conséquence un score moyen entre 1 et 5, que l'on a transformé à nouveau en échelle spam/basse qualité/haute qualité. Nous avons procédé ainsi pour fusionner les notations différentes d'une même page (sinon il est difficile d'arbitrer sur une page qui est du spam selon certains, et de haute qualité selon les autres).

Les résultats préliminaires présentés ici sont issus de l'application de ce procédé sur 10 430 pages.

Quelques résultats intéressants pour le SEO

Nous n'allons pas faire ici le listing exhaustif de tout ce que nous avons analysé, mais plutôt voir certains points intéressants pour le SEO. Dans quelques mois, nous rendrons public une liste exhaustive de tous les critères avec une analyse fine.

Domaine de premier niveau (TLD : Top-Level Domain)

C'est la première question que l'on se pose lorsqu'on veut lancer un nouveau site web : quel TLD faut-il choisir pour inspirer confiance au moteur ? Le problème est par exemple évoqué chez Moz [3].

La figure 2 présente les TLD pour les pages de haute qualité. Le lecteur attentif remarquera la présence de « gouv.fr » qui est un root domain et pas un TLD, mais nous l'avons pris en compte car les sites en sous-domaine de gouv.fr sont généralement indépendants. Sans surprise, il existe une plus forte proportion de sites de haute qualité en gouv.fr et en asso.fr (par exemple, plus de 50% des sites gouv.fr sont de haute qualité). Concernant les autres TLD, on ne trouve pas d'autres spécificités.

Concernant les pages de qualité médiocre (figure 3), rien à signaler de particulier sauf un léger biais au niveau des .info, qui sont 80% du temps des pages de basse qualité.

Concernant le spam, nous n'avons pas identifié énormément de surprises. Notre dataset pour les résultats préliminaires est « tombé » sur blogspot, et donc on voit que la plateforme est utilisé quasiment uniquement pour y mettre du spam. On voit également que les .tv sont plus de 40% du temps du webspam. Ce qu'il faut remarquer, c'est l'absence (par exemple) des .info, on verra sur d'autres critères le même phénomène : les spammeurs sont volontaires dans leur démarche et essayent d'éviter de créer des footprints « spammy ».

La morale de ce premier résultat est que le .org reste toujours le meilleur compromis en terme d'annonce de qualité : pas plus spammé que les autres, et en moyenne les pages en .org sont plus haute qualité (exception faite des TLD moins accessible).

La triplette du bourrin

Pour ceux qui ne connaissent pas cette pratique, il s'agit de mettre les mots-clés visés dans le triplet URL, title et H1. Sur ce critère, on voit le fameux phénomène que nous évoquions plus haut.

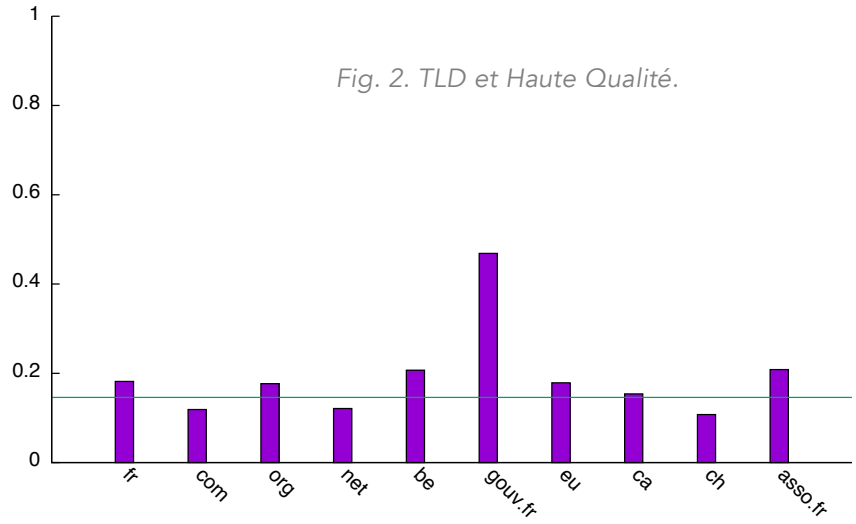

FIGURE: 10 TLD les plus représentés dans les pages de Haute Qualité

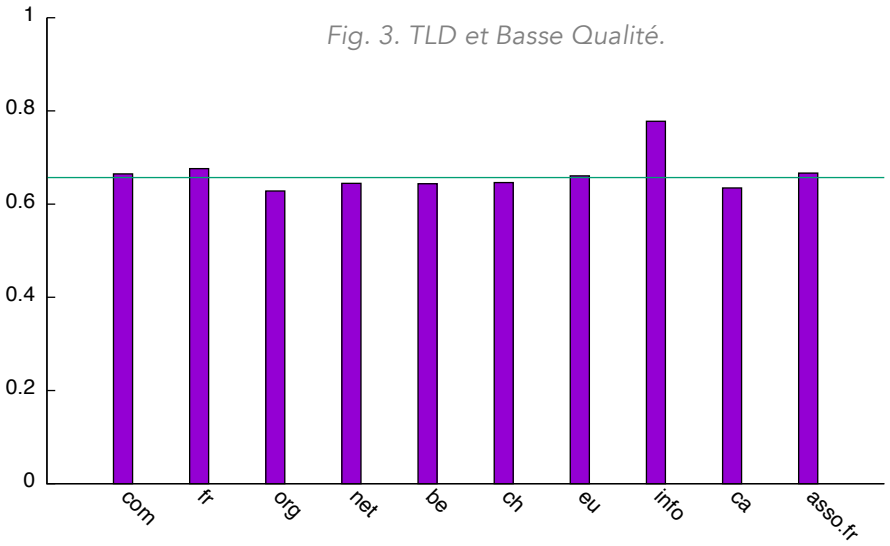

FIGURE: 10 TLD les plus représentés dans les pages de Basse Qualité

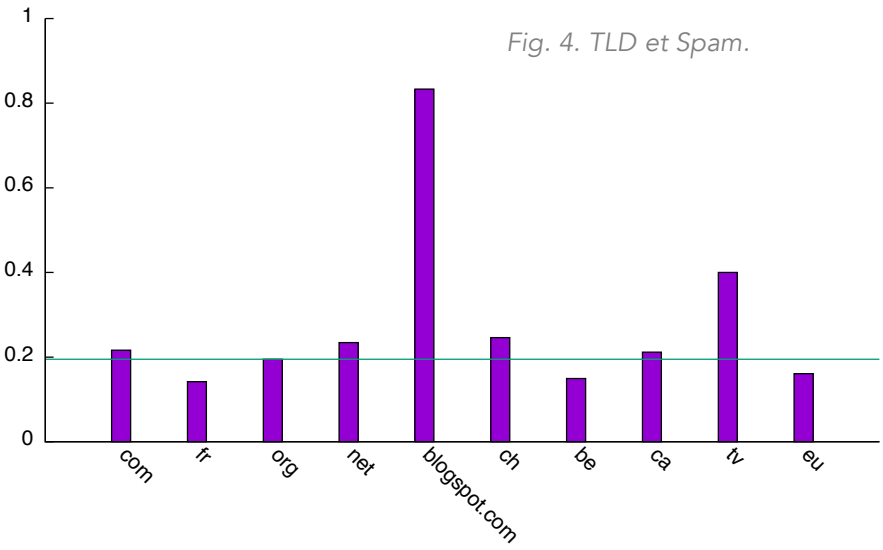

FIGURE: 10 TLD les plus représentés dans les pages SPAM

Comment lire la figure 5 ? En ordonnées se trouve le pourcentage de pages qui ont telle ou telle qualité, et en abscisses on trouve la valeur minimal du critère pour les pages. Ainsi, les trois barres au dessus du « 0 » correspondent aux pages qui ont une distance de Jaccard d'au moins 0 entre URL, title et H1. Les trois barres au dessus du « 0,2 » correspondent aux pages avec un distance de Jaccard de 0,2 ou plus.

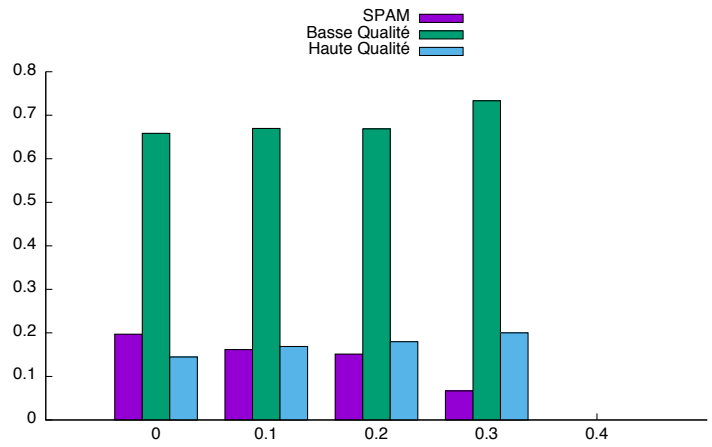

FIGURE: Qualité par rapport à la distance de Jaccard entre URL, title et h1

Fig. 5. Triplette du bourrin

Les barres au dessus du 0 sont donc celles qui caractérisent l'intégralité du dataset. Enfin, plus la distance de Jaccard est grande (elle vaut au plus 1) plus URL, H1 et title contiennent exactement les mêmes mots.

Les résultats numériques sont édifiants : les webmasters qui font des sites de faible ou haute qualité pratiquent (de manière consciente ou inconsciente ? Nous n'en savons rien) la triplette du bourrin. Il est probable que cela soit dû au fait que la plupart des webmasters ne sont pas sensibilisés au SEO.

En revanche, les spammeurs ont peur d'être pénalisés pour optimisation "triple", et ils font donc très attention à ne pas utiliser les mêmes mots dans les trois blocs sémantiques. On voit donc une vraie chute de la proportion de pages de spam quand la distance de Jaccard passe de 0,2 à 0,3 (de 15% à environ 7%, donc une baisse de 50%), cela signifie que dès qu'on dépasse un certain seuil d'optimisation de la triplette, il y a moitié moins de spam.

Taille du title

Sur ce critère, l'information que nous donne notre étude est sans aucune surprise : plus la taille du title est grande (en nombre de caractères) plus la probabilité que la page en question soit du spam ou de mauvaise qualité est grande. Notamment, au-delà d'une taille entre 200 et 300 caractères il n'y a plus que du spam ou de la basse qualité. C'est un des critères qui est finalement assez paradoxal : on pourrait croire que les webmasters qui ont des pratiques borderline seraient sensibilisés au fait que

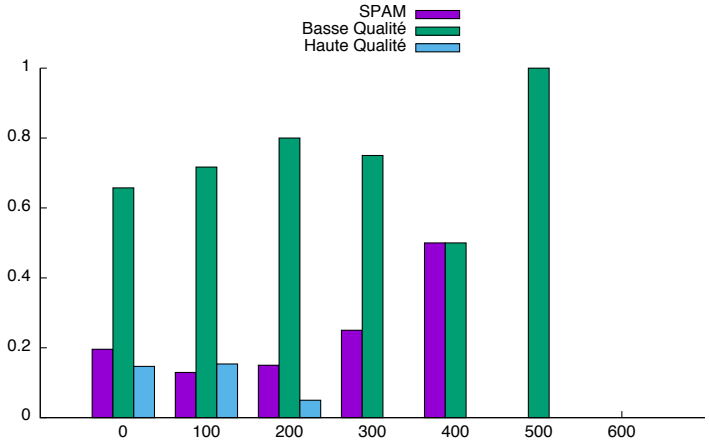

FIGURE: Qualité par rapport à la taille du title

Fig. 6. Taille du Title

maximiser la taille du title est un signal simple et clair pour les moteurs.

Il faut donc essayer de garder ce paramètre dans des tailles raisonnables (autour de 100 caractères par exemple).

Ratio entre taille du footer et taille de la page

Le folklore SEO est assez consensuel sur ce sujet : il faut éviter d'avoir des footer de taille trop importante car 1) cela ne sert à rien et 2) le moteur s'en sert comme critère de pénalisation. Sans rentrer dans cette discussion, nous pouvons déjà étudier la corrélation entre la qualité perçue et la taille du footer. Plus précisément, nous allons étudier le ratio entre la taille du footer et la taille de la page qui contient ce footer.

La figure 7 présente les valeurs numériques associées à ce critère pour notre dataset. Une valeur de 0,05 (comme celle de la deuxième colonne) signifie que le footer représente 5% de la page qui le porte.

On remarque immédiatement que les pages de très haute qualité ont des footers qui sont de proportion raisonnable : principalement en dessous de 10%. Au delà de cette valeur, on trouve une très nette augmentation de la présence des pages qui ont été notées spam, avec une inversion à partir de 25%, seuil au delà duquel le spam est égalitaire ou majoritaire face aux pages de basse qualité.

Il paraît donc raisonnable d'être prudent sur le footer que l'on va créer sur son site web. D'autant plus que les liens sitewide en footer ont désormais un impact très faible au niveau du positionnement.

Les partages Facebook

Autre mythe que l'on croise au détour des forums et blogs : celui des partages sociaux. La question que l'on va considérer ici est de savoir si partages et qualité vont de pair. On pourrait effectivement croire qu'un contenu de qualité est plus partagé, et, réciproquement, que si un contenu est très partagé, c'est sans doute qu'il est de qualité.

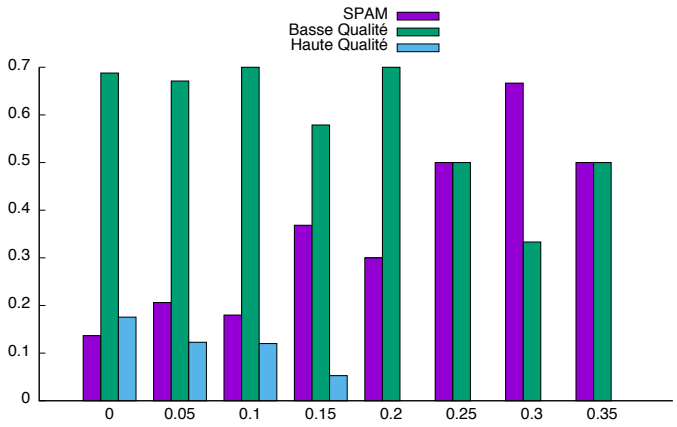

FIGURE: Qualité versus ratio footer/total

Fig. 7. Ratio footer / taille de la page

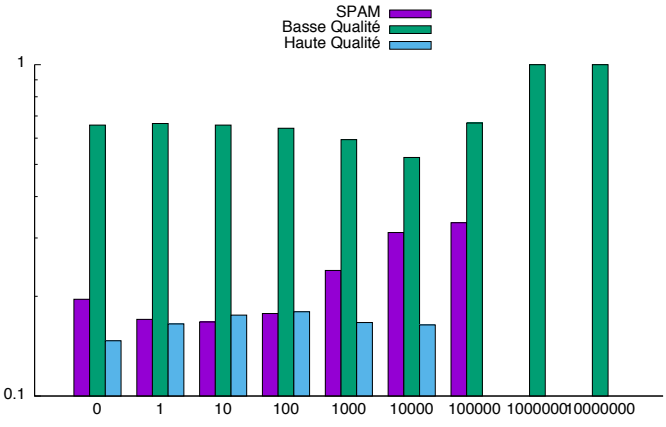

FIGURE: Qualité versus le nombre de partages Facebook

Fig. 8. Partages Facebook

On pourrait le croire, mais c'est faux. Si l'on fait abstraction des nombres énormes et donc ridicule car correspondant à des singularités ou à des manipulations éhontées, on s'aperçoit que la proportion de contenus de haute qualité est sensiblement la même pour des pages non partagées, avec 10, 100, 1 000 voire 10 000 partages. En revanche, au delà d'un seuil qui est de l'ordre de la centaine de milliers de partages, on ne trouve plus que du spam, puis de la basse qualité (du contenu conçu pour hameçonner les internautes).

Conclusion (et la suite)

Dans cet article, nous avons vu les premiers résultats de l'étude webspam. L'étude se poursuit, il faut plus de notations (direction le site web http://webspam.peyronnet.eu pour nous aider) et nous sommes en cours d'analyse de nombreux autres critères.

Par la suite, ces critères seront regroupés dans des systèmes de règles pour fournir des listes de bonnes pratiques aux SEOs. Il s'agit donc d'une affaire à suivre !

Références

[1] Ntoulas, A., Najork, M., Manasse, M., & Fetterly, D. (2006, May). *Detecting spam web pages through content analysis*. In Proceedings of the 15th international conference on World Wide Web (pp. 83-92). ACM. http://www.ra.ethz.ch/CDStore/www2006/devel-www2006.ecs.soton.ac.uk/programme/files/pdf/3052.pdf

[2] Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.

[3] Choix du domaine. https://moz.com/learn/seo/domain

Guillaume Peyronnet *est gérant de Nalrem Médias.* **Sylvain Peyronnet** *est co-fondateur et responsable des ix-labs, un laboratoire de recherche privé. Ensemble, ils font des formations, pour en savoir plus : http://www.peyronnet.eu/blog/*